Unified Evaluation Of Neural Network Calibration & Refinement

Aditya Singh, Alessandro Bay and Andrea Mirabile

Zebra Technologies, London, United Kingdom

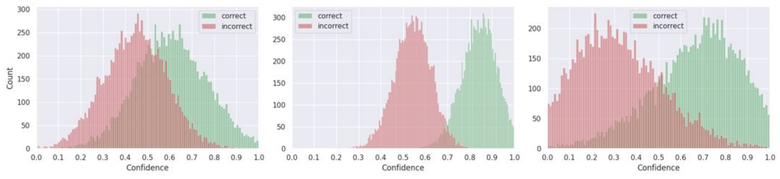
Introduction

Modern neural networks are highly uncalibrated. It poses a significant challenge for safety-critical systems to utilise deep neural networks (DNNs), reliably. Many recently proposed approaches have demonstrated substantial progress in improving DNN calibration. However, they hardly touch upon refinement, which historically has been an essential aspect of calibration. This paper presents a theoretically and empirically supported exposition for reviewing a model's calibration and refinement. We show the breakdown of expected calibration error (ECE), into predicted confidence and refinement. We show through empirical evaluations of many state of the art calibration approaches on standard datasets that many calibration approaches with the likes of label smoothing, mixup etc. lower the utility of a DNN by degrading its refinement.

Related Work

- Refinement and calibration have often been studied together in the field of statistics(Gneiting, Balabdaoui, and Raftery 2007), meteorology(Murphy and Winkler 1977), medicine(Tversky and Kahneman 1974) etc.
- Some examples of existing calibration methods are ERL(Pereyra et al. 2017), LS(Müller, Kornblith, and Hinton 2019), MX(Thulasidasan et al. 2019).
- Approaches discussing are but not limited to CFN(Corbière et al. 2019), CRL(Moon et al. 2020).
- Metrics used for measuring calibration are ECE, OE(Naeini, Cooper, and Hauskrecht 2015), Brier score(Brier 1950).
- Refinement metrics are AUROC, AUPR, FPR@95%-TPR.

Calibration & Refinement



The figures above displays hypothetical classification scenario for 3 classifiers(50% accuracy for each). In (a), we have a case where calibration is good but the refinement is poor. In (b), refinement is good but the calibration is bad. In (c), we have a scenario where calibration and refinement both are good. (c) is the ideal case which we wish to achieve.

References

Brier, G. W. (1950). "Verification of forecasts expressed in terms of probability". In: *Monthly weather review*.

Corbière, Charles et al. (2019). "Addressing Failure Prediction by Learning Model Confidence". In: *NeurIPS*.

Gneiting, Tilmann, Fadoua Balabdaoui, and Adrian E. Raftery (2007). "Probabilistic forecasts, calibration and sharpness". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Hernández-Orallo, José, Peter Flach, and Cèsar Ferri (2012). "A Unified View of Performance Metrics: Translating Threshold Choice into Expected Classification Loss". In:

Moon, Jooyoung et al. (2020). "Confidence-Aware Learning for Deep Neural Networks". In: *ICML*.

Müller, Rafael, Simon Kornblith, and Geoffrey E Hinton (2019). "When does label smoothing help?" In: *NeurIPS*.

Murphy, Allan H. and Robert L. Winkler (1977). "Reliability of Subjective Probability Forecasts of Precipitation and Temperature". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)*.

Naeini, Mahdi Pakdaman, Gregory F. Cooper, and Milos Hauskrecht (2015). "Obtaining Well Calibrated Probabilities Using Bayesian Binning". In: *AAAI*.

Pereyra, Gabriel et al. (2017). "Regularizing Neural Networks by Penalizing Confident Output Distributions". In: *ICLR, Workshop Track Proceedings*.

Thulasidasan, S. et al. (2019). "On Mixup Training: Improved Calibration and Predictive Uncertainty for Deep Neural Networks". In: *NeurIPS*.

Tversky, Amos and Daniel Kahneman (1974). "Judgment under Uncertainty: Heuristics and Biases". In: *Science*.

ECE & AUROC

ECE is a calibration metric widely used in measuring calibration of a model. It is defined as:

$$ECE = \sum_{m} \frac{|B_m|}{n} [|\mathbb{E}[A]_m - C_m|] \tag{1}$$

where, average confidence (C_m) and accuracy (A_m) is computed after splitting the predictions in to predefined m bins based on the predicted confidence and n are the total number of predicted samples.

Assumptions: We employ the following assumptions

- The number of bins, $m = 1$, for computing ECE. Though, this assumption can be relaxed and the final derivation will still hold.
- $\mathbb{E}[A] < C$. Model is over-confident and less accurate. This scenario describes the problem of modern DNNs.

AUROC captures the concept of ordinal ranking nicely. It denotes the expectation that a uniformly drawn random positive is ranked higher than a uniformly drawn random negative sample. AUROC (r), is formally defined as

$$r = \int_0^1 tpr \; \mathrm{d}fpr, \tag{2}$$

where, tpr is the true positive rate and fpr is the false positive rate. The maximum value that r can attain is 1 representing an ideal ranking scenario.

We build on the work of (Hernández-Orallo, Flach, and Ferri 2012) to show that,

$$ECE = \alpha C - \beta r - \gamma, \tag{3}$$

where $\alpha \geq \beta > 0$ and $\gamma \geq 0$.

Results

We utilize LS:label smoothing, MX:mixup, ERL:entropy regularisation as calibration methods. As refinement methods, we employ CRL:correctness ranking loss and CFN:confidence network. We perform the joint evaluation on CIFARs and STL-10 datasets. The DNN we use is the VGG-16 with batch normalisation.

Table 1: CIFAR-100

Method	Accuracy(↑)	ECE(↓)	AUROC(↑)
Baseline	72.07 ± 0.2	19.12 ± 0.13	85.18 ± 0.21
ERL	72.40 ± 0.19	16.8 ± 0.1	85.19 ± 0.3
Mixup	73.12 ± 0.18	6.87 ± 1.81	82.96 ± 0.27
LS	72.92 ± 0.43	**5.76 ± 0.56**	81.49 ± 0.27
CFN	72.07 ± 0.2	13.95 ± 2.7	86.0 ± 0.18
CRL	71.5± 0.2	12.5 ± 1.1	**88.11 ± 0.16**

Table 2: CIFAR-10

Method	Accuracy(↑)	ECE(↓)	AUROC(↑)
Baseline	92.96 ± 0.2	5.38 ± 0.15	92.5 ± 0.01
ERL	93.23 ± 0.01	4.41 ± 0.07	92.11 ± 0.4
Mixup	**93.46 ± 0.18**	4.16 ± 1.2	86.72 ± 0.8
LS	93.07 ± 0.2	7.4 ± 0.18	82.36 ± 1.23
CFN	92.96 ± 0.2	4.1 ± 0.2	92.55 ± 0.1
CRL	93.05 ± 0.37	**1.87 ± 0.21**	**92.59 ± 0.42**

Results

Table 3: STL-10

Method	Accuracy(↑)	ECE(↓)	AUROC(↑)
Baseline	81.61 ± 0.2	11.55 ± 0.19	85.53 ± 0.7
ERL	82.38 ± 0.29	9.6 ± 0.41	86.57 ± 0.16
Mixup	**82.94 ± 0.08**	**3.46 ± 0.33**	85.9 ± 0.14
LS	81.99 ± 0.45	5.64 ± 0.02	85.13 ± 0.3
CFN	81.61 ± 0.2	9.23 ± 1.02	**86.64 ± 0.4**
CRL	79.5 ± 0.4	6.34 ± 1.19	85.29 ± 0.68

Observations

- Calibration approaches provide better calibration and refinement based approaches provide better refinement.
- Calibration based approaches perform poorly in comparison to the uncalibrated baseline in terms of refinement. This highlights a significant drawback in these approaches which has not yet been highlighted.
- Refinement based approaches provide good calibration. This provides empirical evidence in support of the relationship derived between ECE and AUROC.

Summary

- We have highlighted the connection between Expected Calibration Error, a calibration metric, and area under the ROC curve computed for a classification task. This result forms the motivation for our cross-domain evaluation. Based on the derived relationship, we discuss the cases where methods focusing on one task can positively or negatively impact the other.
- We evaluate respective state of the art methods which are studied in isolation under a unified setting and showed that refinement based approaches improve calibration.
- We also showed that calibration based approaches perform poorly on refinement.

Future Work

- Understanding the reasoning behind the drop in refinement for calibration methods.
- Extend evaluation to more deep neural network architectures and datasets.
- Device refinement based approach better at calibration than the existing calibration approaches.
- Study the refinement-calibration trade-off in scenarios where there is data shift.